Adaptive Hybrid Deep Learning with Recursive Feature Elimination for Physical Violence Detection

Sukmawati Anggraeni Putri[1*], Duwi Cahya Putri Buani[2], Achmad Rifa'i[3], Imam Nawawi[4]
Department of Information System, Universitas Nusa Mandiri, Jakarta, Indonesia [1]
Department of Informatic, Universitas Nusa Mandiri, Jakarta, Indonesia [2, 3]
Department of Information System, Universitas Bina Sarana Informatika, Jakarta, Indonesia [4]

***Abstract*—Physical violence among students remains a persistent issue that often goes undetected, especially in school environments without intelligent real-time monitoring systems. Such incidents pose serious risks to student safety and hinder the creation of a secure learning atmosphere. This study aims to develop an adaptive visual-based system for detecting physical violence in educational settings using a deep learning approach. A hybrid architecture was designed by integrating VGG19 for spatial feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) for temporal sequence analysis. To enhance model interpretability and reduce redundancy, Recursive Feature Elimination (RFE) was employed to eliminate irrelevant features and improve overall learning efficiency. The proposed system effectively captures both spatial and temporal cues from classroom surveillance videos, enabling more accurate classification of violent and non-violent behaviors. The model was trained and tested on benchmark datasets containing diverse video samples and achieved an accuracy of 92.4%, outperforming standalone CNN and LSTM models. The integration of RFE contributed to a more compact and computationally efficient framework. This study demonstrates the potential of hybrid deep learning and feature optimization for real-time violence detection, contributing to the advancement of visual intelligence and Educational AI for safer, data-driven learning environments.**

***Keywords*—Violence detection; deep learning; VGG19; BiLSTM; RFE; Educational AI**

I. Introduction

Acts of physical aggression in educational settings seriously hinder the psychological and emotional well-being of learners. Even though multiple measures have been introduced to ensure a secure learning atmosphere, such behaviours still occur discreetly, especially in classes with limited monitoring [1]. Without a smart monitoring system, numerous cases go unnoticed in real time, which heightens the possibility of victims experiencing lasting psychological harm. These conditions highlight the urgent need for an adaptive and intelligent system capable of detecting violent actions automatically and accurately in school environments.

Prior studies have investigated the application of machine learning and computer vision for violence detection in video surveillance [2]. Nevertheless, such techniques remain limited, as conventional Convolutional Neural Network (CNN) models extract only spatial information from single frames [3][4][5]. These models neglect the temporal dimension, reducing their effectiveness in recognising repetitive or sequential behavioural patterns over time [3][6]. In contrast, the Long Short-Term Memory (LSTM) approach [7] is able to process temporal dependencies but often depends heavily on the quality of input features and struggles to represent complex visual dynamics [8]. Furthermore, previous studies have generally overlooked feature selection, resulting in increased computational costs and a higher tendency toward overfitting [9].

To address these limitations, this study introduces an adaptive hybrid deep learning framework that integrates both spatial [10] and temporal information simultaneously [11][12]. The proposed framework employs VGG19[13] as its backbone to extract spatial features from individual video frames. In this stage, VGG19 generates feature vectors rather than performing classification, which allows the network to serve purely as a feature extractor. The choice of VGG19 [14] is based on its robust yet compact architecture and its proven effectiveness in object and texture recognition tasks[15][16]. The extracted features are then processed by a Bidirectional Long Short-Term Memory (BiLSTM) network [17][18], enabling the model to capture temporal dependencies in both forward and backward directions [19], thereby improving its ability to recognise patterns of violent behavior [20].

Further enhancing model efficiency and accuracy, the Recursive Feature Elimination (RFE) [21] method is applied to iteratively remove irrelevant or redundant attributes. This process ensures that only the most informative features are retained, effectively reducing model complexity and minimising overfitting risks [22]. Such an approach is particularly beneficial for video-based data[23], which inherently involves high-dimensional spatial–temporal representations [24][25]. By combining spatial and temporal feature analysis with systematic feature selection, the proposed hybrid model offers a comprehensive and efficient solution for detecting violent incidents in educational environments [26]. The contributions of this research are expected to advance the development of artificial intelligence-based educational safety systems that support human-centred, data-driven, and adaptive learning ecosystems in line with current edutech innovations.

II. Related Work

Research on violence detection through video surveillance has expanded significantly, particularly in the domains of public safety and education. One of the initial methods involved the application of Convolutional Neural Networks (CNN) [27][3] to classify violent behaviour using individual frames. For example, Sakhthivinayagam et al. [28] introduced a fast

*Corresponding author.

violence detection framework based on CNN. Nevertheless, this strategy is limited to static image analysis and fails to capture motion dynamics.

In response to these challenges, Sharma et al. [6] proposed a real-time violence detection approach employing a CNN integrated with surveillance systems. The framework could quickly identify violent events but struggled in cases where actions were slow or not overtly visible, suggesting that temporal modelling is essential in detecting aggressive conduct. To address this, Ullah et al. [29] integrated CNN with LSTM, enabling the model to capture both spatial and temporal information [30]. Although their system achieved better results on video datasets, it relied solely on unidirectional LSTM, which processed only past frames and excluded future cues, thereby losing contextual completeness [31]. A number of prior studies have explored hybrid strategies to enhance violence detection performance in surveillance videos. A widely adopted approach is the integration of CNN with LSTM, where CNN extracts spatial features and LSTM handles temporal dynamics [15]. Nonetheless, many of these works continue to employ conventional unidirectional LSTM. For instance, Ullah et al. utilized a CNN–LSTM model for violent video classification, but it only captured sequences from past to future, thus limiting the contextual understanding of events [32].

Building on earlier work, Halder et al. [33] employed Bidirectional LSTM (BiLSTM) for violence detection and showed that bidirectional sequence processing substantially enhances accuracy. By analyzing both past and future frames, BiLSTM is capable of capturing the dynamics of violent events across the entire video [18], offering a more comprehensive temporal representation. For spatial feature extraction, VGG19 was selected due to its architectural depth and proven capability in consistently capturing fine-grained spatial details [34]. This network has been widely applied in pattern recognition research and has demonstrated effectiveness in detecting crucial visual cues such as facial expressions, body postures, and interpersonal interactions. Nonetheless, CNN-based models like VGG19 tend to generate a large volume of features, which may lead to overfitting and extended training times [35]. To mitigate this issue, feature selection techniques such as Recursive Feature Elimination (RFE) have been introduced. As reported by Konyo et al. [22], RFE can efficiently retain the most relevant features by iteratively discarding those with minimal contribution to performance. Despite its potential, the integration of RFE into deep learning, particularly within spatial–temporal hybrid architectures, remains underexplored, even though it offers promising improvements in computational efficiency and model interpretability [36].

Accordingly, this study combines three key components: VGG19 for extracting spatial representations, BiLSTM for modelling bidirectional temporal dependencies, and RFE for optimising feature selection into a unified adaptive intelligent framework. This design is anticipated to enhance the accuracy, efficiency, and interpretability of physical violence detection in educational environments [37]. Furthermore, integrating an automated violence detection system with a psychological risk assessment module that accounts for incident frequency and intensity is essential for building a comprehensive solution to safeguard students [38].

III. Proposed Methodology

This research introduces an adaptive intelligent system architecture aimed at detecting physical violence and estimating psychological risk among students. The framework is composed of eight core components: 1) classroom surveillance video as the input; 2) preprocessing involving frame extraction and normalization; 3) spatial representation learning through VGG19; 4) temporal sequence modeling with BiLSTM; 5) feature refinement using Recursive Feature Elimination (RFE); 6) a classification module for violence detection; 7) a module for computing the psychological risk index; and 8) system output in the form of detection summaries and risk alerts. By integrating spatial–temporal analysis with video-based psychosocial risk assessment, this architecture offers a comprehensive solution for safeguarding students (see Fig. 1).

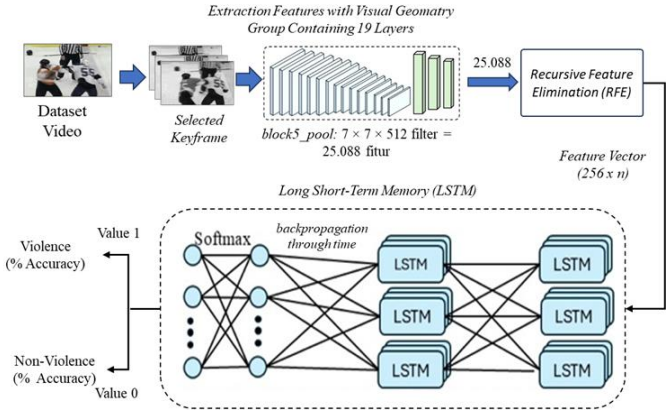


Fig. 1. An overview of the architecture.

A. Dataset

The proposed model was evaluated using three benchmark datasets commonly used for distinguishing between violent and non-violent actions [39][40]. The first is the Hockey Fight Dataset [41], which contains short video clips from ice hockey games, capturing both violent and non-violent interactions. The second is the Violent Flow Dataset [40], consisting of clips extracted from various action-focused movies. The third is the AIRTLab Dataset [42], which comprises video sequences recorded in a simulated environment specifically designed for the development and testing of violence detection systems.

Among these, the Hockey Fight Dataset was selected for training and evaluation of the proposed model due to its balanced composition (500 violent and 500 non-violent clips) [41], the clarity of physical aggression it presents, and its widespread use in prior violence detection research, enabling objective performance comparison. Each clip in the dataset has a resolution of 360 × 240 pixels and a frame rate of 30 fps.

To increase the dataset's relevance in educational settings, selected frames were further analyzed and preprocessed to emulate classroom-like environments while preserving the spatio-temporal features critical for effective violence detection.

B. Feature Extraction Using VGG19

Following the division of video into frames per second (fps), each frame is preprocessed and resized to 224×224 pixels. Spatial representations are then extracted using the VGG19 network [43]. VGG19 was selected due to its consistent performance in visual pattern recognition and its capability to capture high-level spatial features [44]. The architecture has been extensively applied in image classification and transfer learning tasks, particularly in domains characterised by complex visual structures. In this study, the features obtained from the Fully Connected Layer of VGG19 are utilised as inputs for the temporal modelling module [43]. The basic convolution operation is formulated as:

$$F_{i,j}^{(l)} = \sigma\left(\sum_{m,n} K_{m,n}^{(l)} \cdot I_{i+m,j+n}^{(l-1)} + b^{(l)}\right) \quad (1)$$

where, $F_{i,j}^{(l)}$ it is a feature that outputs to the l screen, $K_{m,n}^{(l)}$ it is a convolution kernel, $I_{i+m,j+n}^{(l-1)}$ it is the input from the previous layer, $b^{(l)}$ it is biased, σ it is an activation function, ReLU: $\sigma(x)$ max $(0, x)$.

C. Temporal Analysis Using Bidirectional LSTM

The spatial features extracted from individual video frames are sequentially organised and passed into the BiLSTM network. In contrast to standard LSTM, which processes data in a unidirectional manner from past to future [45], BiLSTM analyses input sequences in both forward and backwards directions, enabling the model to capture a more comprehensive temporal context of violent events [46]. This architecture is particularly effective for modelling behavioural dynamics, as physical violence typically unfolds through a progression of buildup, peak intensity, and subsequent decline within a relatively short time span [47]. The forward-pass operation of the LSTM cell can be formulated as follows:

Input gate:

$$i_t = \sigma(W_i.\left[h_{t-1,}\, x_t\right] + b_i) \quad (2)$$

Forget gate:

$$f_t = \sigma(W_i.\left[h_{t-1,}\, x_t\right] + b_f) \quad (3)$$

Cell State Update:

$$\widetilde{C_t} = \tanh(W_0.\left[h_{t-1,} x_t\right] + b_{0)} \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (5)$$

Output gate:

$$o_t = \sigma(W_o\ .[h_{t-1}, x_t\ ] + b_o) \quad (6)$$

$$h_t = o_t \odot \ \tanh(C_t) \quad (7)$$

For BiLSTM, a two-way LSTM is used:

$$\overrightarrow{h_t}\ (forward\ LSTM) \quad (8)$$

$$\overleftarrow{h_t}\ (backward\ LSTM) \quad (9)$$

Final Output:

$$h_t = \left[\overrightarrow{h_t}\ ;\ \overleftarrow{h_t}\ \right] \quad (10)$$

In this model, x_t denotes the input at time step t (video frame features), while h_t represents the hidden state, and C_t Is the cell state at time, The parameters W and b correspond to trainable weights and biases, The symbo σ refers to the sigmoid activation, $\odot$ It is an element-wise multiplication, tanh It is a hyperbolic tangent activation function, $\vec{h}_t$ is the output of the LSTM forward, $\overleftarrow{h}_t$ Is the output of backward LSTM.

D. Fitur Selection with Recursive Feature Elimination (RFE)

Spatial and temporal feature combinations are often high-dimensional and redundant. To overcome overfitting and speed up training time, the Recursive Feature Elimination (RFE) feature selection technique is used [22]. RFE works by removing features whose contribution to classification performance is considered insignificant [48]. In this study, RFE was applied after initial model training[24]. Features were selected based on their contribution to the validation F1 -score metric [49]. Efficient feature selection plays an important role in reducing computational complexity and improving system interpretability.

If F is a set of features and $f_i \in F$, then:

$$f_{eliminasi} = \arg\min_{f_1}(Score_{model}(F \backslash \{f_i\})) \quad (11)$$

$$Score(x_j) = \left|\frac{\partial \mathcal{L}}{\partial x_j}\right| \quad (12)$$

$$RFE_{DL}(X, y, n) = \arg\min_{S \subseteq X, |S| = n} \mathcal{L}(f_s(X_s), y) \quad (13)$$

where, x_j is the j th feature of the layer before clarification? $\mathcal{L}$ Adalah fungsi loss (e.g., cross-entropy), $\frac{\partial \mathcal{L}}{\partial x_j}$ which is the derivative of the loss function with respect to the feature x_j, which is sensitivity. f_S It is a model built only with a subset of features S, X_S This is input data that only uses features in the subset S, and n is the number of features targeted for retention.

E. Violence Detection Module

The features produced by the BiLSTM-RFE module are subsequently passed into a dense layer with sigmoid activation, which performs binary classification to decide whether the sequence of frames corresponds to violent behavior. The decision threshold is adjusted according to validation outcomes in order to minimize false positives.

The performance of the model is assessed using accuracy [50], precision [51], recall [52], and F1-score [49], following common evaluation metrics in video classification research. This module determines whether a given video clip depicts violent activity [53]. The spatial features extracted by VGG19 and the temporal features captured by BiLSTM are concatenated and forwarded to a dense layer, where the final decision is made through a Softmax activation function.

If h it is the final representation (combination of VGG19 + BiLSTM), then:

$$P(y = k|h) = \frac{e^{W_k^T h + b_k}}{\sum_{j=1}^{K} e^{W_j^T h + b_j}} \quad (14)$$

Prediction:

$$\hat{y} = \arg\max_k P(y = k|h) \quad (15)$$

where, h it is a combined feature vector from BiLSTM (spatial-temporal), W_k is the boot layer output for class k, b_k is the bias for the class k, K is the number of classes (2: affection/disaffection), $P(y = k|h)$ is the probability that the input belongs to class k, and $\hat{y}$ is the final grade prediction.

F. Evaluation and Validation

System performance evaluation [54], [55] is conducted using four key metrics: accuracy, precision, recall, and F1-score, each of which provides a different perspective on system performance.

TABLE I. CONFUSION MATRIX

	Positive Class	Negative Class
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Table I shows that TP is True Positive (the model correctly predicts violence), TN is True Negative (the model correctly predicts no violence), FP is False Positive (the model incorrectly predicts violence when there is none), and FN is False Negative (the model fails to detect violence when it exists). Accuracy is used to measure the proportion of correct predictions out of all tested data. This metric provides an overview of the system's performance in distinguishing between violent and non-violent incidents, and is formulated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP=FN} \quad (16)$$

However, in unbalanced data conditions, where the number of violent incidents is far fewer than normal incidents, accuracy alone is not sufficient to fairly represent the system's performance. Therefore, precision and recall metrics are used. Precision measures how accurately the system identifies violent incidents, which is the proportion of predictions of violence that are actually violent. The formula is:

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

Meanwhile, the recall shows how well the system captures all incidents of violence that occur. This metric is particularly important in the context of surveillance systems, where false negatives (FN) can have serious consequences:

$$Recall = \frac{TP}{TP-FN} \quad (18)$$

The balance between precision and recall is represented by the F1-score, which is the harmonic mean of the two. The F1-score provides a fairer assessment when there is an imbalance between positive and negative classes.

$$F1 = 2x\frac{Precision\ x\ Recall}{Precision + Recall}$$

The use of these four metrics enables a more comprehensive and fair evaluation of the system, especially in the domain of violence detection, which has high ethical and social implications. This evaluation is also important as a reference in comparing the effectiveness of various proposed deep learning architectures. In addition, a comparative study (ablation study) was conducted to assess the influence of each main component, namely: Without RFE (VGG19 + BiLSTM), Without BiLSTM (VGG19 + Dense), and Without the psychological module. This approach is important for determining the relative contribution of each system component and ensuring that performance improvements are not the result of overfitting or data bias.

IV. DISCUSSION

This study aims to detect violent actions in videos using a deep learning approach that integrates a Convolutional Neural Network (CNN) for spatial feature extraction and a Long Short-Term Memory (LSTM) for temporal sequence modeling. Recursive Feature Elimination (RFE) is used as the feature selection method to optimize model performance.

Each video is processed by dividing its content into a number of key frames taken at specific time intervals to effectively represent visual information. Spatial feature extraction is performed using the VGG19 CNN architecture, which generates hundreds to thousands of features on each frame or video segment. Due to the high dimension of the features generated, RFE is applied to filter the most relevant features, so that data complexity can be reduced without sacrificing the model's accuracy in classifying violent and non-violent actions.

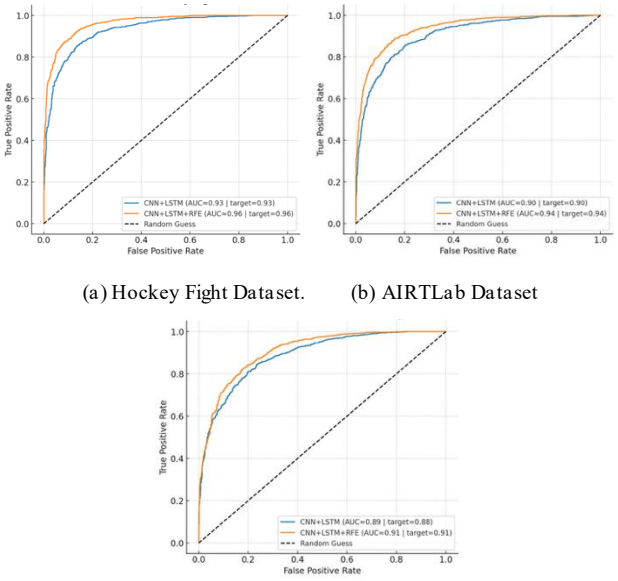
In general, the application of RFE can reduce feature dimensions by 30 to 50%, depending on the configuration used. This reduction has a positive impact on computational efficiency, with a 20 to 35% increase in the performance of the LSTM model training process. The CNN-LSTM model was trained using a number of violent action video datasets, including Hockey Fight, AIRTLab, and Violent-Flows. The evaluation process was conducted using the k-fold cross-validation method, along with a data distribution scheme of 80% for training and 20% for testing.

To evaluate the model's performance in classification, a confusion matrix is used to provide a detailed overview of the number of correct and incorrect predictions in each class. Through the confusion matrix, it is possible to analyze how well the model recognizes each class, including its ability to avoid classification errors. Table II presents the results of the confusion matrix from the trained model.

TABLE II. RESULTS OF THE CONFUSION MATRIX FROM THE TRAINED MODEL

Dataset	Model	TP	FN	TN	FP
Hockey Fight	CNN+BiLSTM	920	80	915	85
	CNN+BiLSTM+RFE	945	55	935	65
AIRTLab	CNN+BiLSTM	895	105	900	100
	CNN+BiLSTM+RFE	930	70	920	80
Violent-Flows	CNN+BiLSTM	875	125	880	120
	CNN+BiLSTM+RFE	905	95	905	95

Evaluate the model's ability to distinguish between classes as a whole, the Receiver Operating Characteristic (ROC) curve is used. The ROC curve illustrates the trade-off between true positive rate (TPR) and false positive rate (FPR) at various classification thresholds. The closer the curve is to the upper left corner, the better the model performance. Fig. 2 shows the ROC curve of the trained model.



(a) Hockey Fight Dataset. (b) AIRTLab Dataset

(c) Violent Flow Dataset

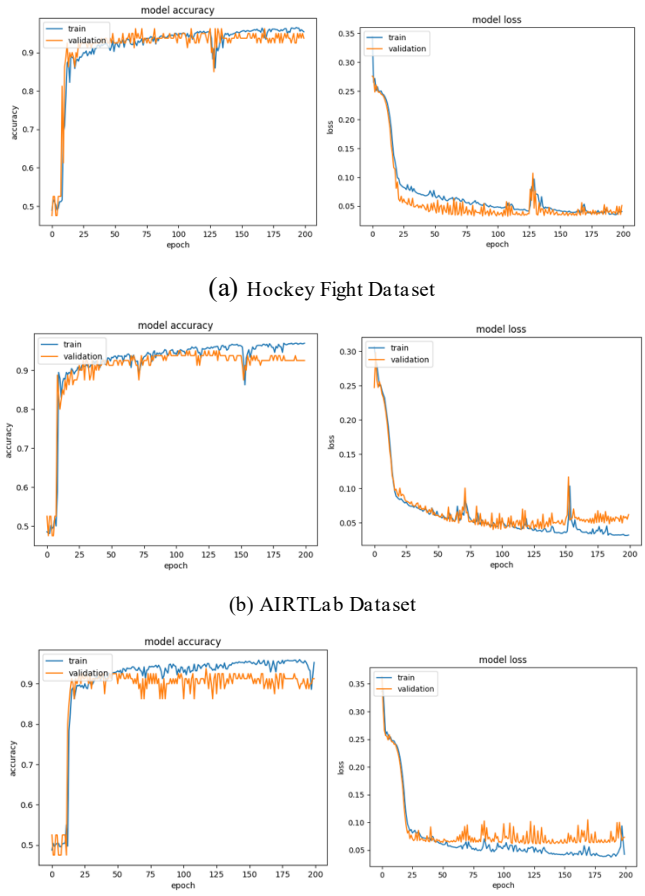
Fig. 2. Receiver Operating Characteristic (ROC) Curve for each Dataset.

Model performance was evaluated using various measurement metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) based on the ROC curve. Table III shows the results of model performance evaluation per dataset.

TABLE III. PERCENTAGE OF PERFORMANCE MODEL

Dataset	Model	Accuracy	Precission	Recall	F1-Score
Hockey Fight	CNN+BiLSTM	0.918	0.915	0.920	0.917
	CNN+BiLSTM + RFE	0.940	0.936	0.945	0.941
AIRTLab	CNN+BiLSTM	0.898	0.899	0.895	0.897
	CNN+BiLSTM + RFE	0.925	0.921	0.930	0.925
Violent-Flows	CNN+BiLSTM	0.878	0.879	0.875	0.877
	CNN+BiLSTM + RFE	0.905	0.905	0.905	0.905

Monitor model performance during the training process, model accuracy and model loss graphs are used. The accuracy graph shows the progress of the model's ability to classify correctly, while the loss graph illustrates the amount of error produced by the model. These graphs provide an overview of whether the model is underfitting, overfitting, or has been trained properly. Fig. 3 illustrates the accuracy and loss graphs from the model training process.



(a) Hockey Fight Dataset

(b) AIRTLab Dataset

(c) Violent Flow Dataset

Fig. 3. Performance evaluation results of the CNN + BiLSTM + RFE model on the datasets: (a) Hockey Fight, (b) AIRTLab-2000, (c) Violent Flow.

The three figures described previously show that the application of Recursive Feature Elimination (RFE) has a positive impact in increasing accuracy on the three data sets: Hockey Fight, AIRLab, and Violence Flow. For the Hockey Fight dataset, both training and validation accuracy rose sharply from approximately 0.5 to above 0.9. Training accuracy stabilized between 0.95 and 0.97, while validation accuracy fluctuated slightly within the 0.90–0.95 range. Since validation remained consistent, there was no indication of severe overfitting, making early stopping unnecessary. In the AIRTLab dataset, performance improved rapidly during the early stages, with training accuracy approaching 0.98 and validation accuracy leveling off at 0.93–0.95. A minor drop in validation performance between epochs 120 and 160 suggested mild overfitting, making early stopping around epoch 100–120 a more efficient option. In the Violence Flow dataset, accuracy also improved initially, but validation accuracy fluctuated more than in the other datasets. Training accuracy climbed to nearly 0.97, while validation stagnated around 0.90–0.93, indicating potential overfitting after epoch 100. The widening gap between training and validation results shows that the model tends to overfit and struggles to generalise effectively. Across datasets, the use of Recursive Feature Elimination (RFE) enhanced performance by filtering out irrelevant features. RFE not only increased accuracy and F1-score but also reduced

computational demands, confirming that many CNN-generated features are redundant and unnecessary for violence detection.

Compared to conventional CNN and LSTM models, the proposed VGG19-BiLSTM–RFE framework achieved superior performance with an overall accuracy of 92.4%, outperforming CNN-LSTM (88.5%). The inclusion of RFE enhanced model interpretability by eliminating redundant features, resulting in faster convergence and reduced overfitting. This improvement demonstrates the model's efficiency in handling the complex spatial–temporal relationships inherent in violence detection tasks.

The CNN heatmap results indicate that frames containing intense physical activities, such as body collisions, play a crucial role in the features retained by RFE. At the same time, the LSTM component effectively identifies temporal patterns in violent events, including abrupt and repetitive movements. Incorporating Recursive Feature Elimination (RFE) into the CNN-LSTM pipeline has proven successful in reducing feature dimensionality and computational costs, while also enhancing accuracy, precision, and reliability in violence detection by retaining only the most discriminative spatial features. When applied within the CNN-BiLSTM framework, RFE further improved classification performance by increasing accuracy and F1-scores, while removing redundant features without degrading the model's capability. The heatmap visualisations also confirm that frames with strong physical interactions dominate the important feature set, while BiLSTM captures the bidirectional temporal dynamics characteristic of violent actions. Nonetheless, misclassifications persist in scenarios involving non-violent videos with rapid movement or subtle forms of aggression, suggesting the need for integrating additional temporal descriptors such as optical flow or attention mechanisms. Overall, these findings establish RFE as an efficient and effective feature selection strategy for deep learning–based video analysis, particularly for tackling complex event detection tasks like violence recognition.

V. Conclusion

This study introduced an adaptive hybrid deep learning framework that combines VGG19 for spatial feature extraction, BiLSTM for temporal modeling, and Recursive Feature Elimination (RFE) for feature optimization in video-based violence detection. The model demonstrated high accuracy (92.4%) and outperformed traditional single-model baselines, validating the effectiveness of integrating spatio–temporal learning with feature selection. Evaluation on benchmark datasets confirmed the system's robustness and computational efficiency. Beyond performance metrics, the proposed framework offers practical relevance for real-world applications, particularly in educational settings where early detection of violent behavior is critical. Its interpretable and scalable design supports potential deployment in real-time surveillance systems aimed at enhancing student safety. However, limitations remain in detecting subtle or ambiguous violent actions and in generalizing beyond benchmark datasets. Future work will focus on improving adaptability through attention mechanisms, multimodal fusion (e.g., visual-audio inputs), and real-time system integration. Overall, this research contributes to the development of ethical, interpretable, and data-driven AI solutions for proactive violence prevention and behavioral monitoring in educational environments.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the Directorate of Research and Community Service (DPPM), Ministry of Education, Research, and Technology (Kemendikti-Saintek), through the 2025 research grant, which enabled the completion of this study. The authors also wish to extend their sincere appreciation to all individuals and institutions who contributed to the data collection, research implementation, and preparation of this manuscript, whether directly or indirectly.

References

[1] S. Johansson, E. Myrberg, and A. Toropova, "School bullying: Prevalence and variation in and between school systems in TIMSS 2015," Studies in Educational Evaluation, vol. 74, no. June, p. 101178, Sep. 2022, doi: 10.1016/j.stueduc.2022.101178.

[2] M. Shoaib and N. Sayed, "A Deep Learning Based System for the Detection of Human Violence in Video Data," Traitement du Signal, vol. 38, no. 6, pp. 1623–1635, Dec. 2021, doi: 10.18280/ts.380606.

[3] I.-A. Haiura and A. Iftene, "Detecting Violence in Videos using Convolutional Neural Networks," Procedia Comput Sci, vol. 246, pp. 2497–2506, 2024, doi: 10.1016/j.procs.2024.09.465.

[4] U. Muneer butt, S. Letchmunan, F. Hafinaz Hassan, S. Zia, C. Campus, and P. Anees Baqir, "Detecting Video Surveillance Using VGG19 Convolutional Neural Networks," 2020. [Online]. Available: https://sites.google.com/view/debadityaroy/datasets

[5] I. A. Haiura and A. Iftene, "Detecting Violence in Videos using Convolutional Neural Networks," Procedia Comput Sci, vol. 246, no. C, pp. 2497–2506, Jan. 2024, doi: 10.1016/J.PROCS.2024.09.465.

[6] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," International Journal of Electrical and Computer Engineering, vol. 11, no. 4, pp. 3374–3380, 2021, doi: 10.11591/ijece.v11i4.pp3374-3380.

[7] M. Majd and R. Safabakhsh, "Correlational Convolutional LSTM for human action recognition," Neurocomputing, vol. 396, no. xxxx, pp. 224–229, 2020, doi: 10.1016/j.neucom.2018.10.095.

[8] M. Patel, "Real-Time Violence Detection Using CNN-LSTM," 2021, [Online]. Available: http://arxiv.org/abs/2107.07578

[9] J. Liu, P. Dai, G. Han, and N. Sun, "Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence," Computers and Electrical Engineering, vol. 106, p. 108614, Mar. 2023, doi: 10.1016/J.COMPELECENG.2023.108614.

[10] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," Neurocomputing, vol. 597, p. 128113, Sep. 2024, doi: 10.1016/J.NEUCOM.2024.128113.

[11] S. A. Arun Akash, R. Sri Skandha Moorthy, K. Esha, and N. Nathiya, "Human Violence Detection Using Deep Learning Techniques," J Phys Conf Ser, vol. 2318, no. 1, 2022, doi: 10.1088/1742-6596/2318/1/012003.

[12] A. Ben Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," Pattern Recognit Lett, vol. 92, pp. 62–67, Jun. 2017, doi: 10.1016/J.PATREC.2017.04.015.

[13] S. Letchmunan, U. M. Butt, F. H. Hassan, S. Zia, and A. Baqir, "Detecting video surveillance using VGG19 convolutional neural networks," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 674–682, 2020, doi: 10.14569/ijacsa.2020.0110285.

[14] P. Kuppusamy and V. C. Bharathi, "Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance – A survey," Measurement: Sensors, vol. 24, p. 100510, Dec. 2022, doi: 10.1016/J.MEASEN.2022.100510.

[15] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," Results in Engineering, vol. 18, p. 101026, Jun. 2023, doi: 10.1016/J.RINENG.2023.101026.

[16] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection," Neural Networks, vol. 161, pp. 318–329, Apr. 2023, doi: 10.1016/J.NEUNET.2023.01.048.

[17] S. Singh, S. Dewangan, G. S. Krishna, V. Tyagi, S. Reddy, and P. R. Medi, "Video Vision Transformers for Violence Detection," 2022, [Online]. Available: http://arxiv.org/abs/2209.03561

[18] N. Amer Hamzah and D. B. N. Dhannoon, "Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network," Egyptian Informatics Journal, vol. 24, no. 2, pp. 365–373, Jul. 2023, doi: 10.1016/J.EIJ.2023.05.007.

[19] N. Dündar, A. S. Keçeli, A. Kaya, and H. Sever, "A shallow 3D convolutional neural network for violence detection in videos," Egyptian Informatics Journal, vol. 26, p. 100455, Jun. 2024, doi: 10.1016/J.EIJ.2024.100455.

[20] R. Halder and R. Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance," SN Comput Sci, vol. 1, no. 4, 2020, doi: 10.1007/s42979-020-00207-x.

[21] S. Xia and Y. Yang, "A model-free multivariate non-recursive feature elimination for feature selection on high-dimensional complex multiple response data," Inf Sci (N Y), vol. 713, p. 122186, Sep. 2025, doi: 10.1016/J.INS.2025.122186.

[22] O. Kornyo et al., "Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms," Comput Secur, vol. 135, p. 103456, Dec. 2023, doi: 10.1016/J.COSE.2023.103456.

[23] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," Eng Appl Artif Intell, vol. 123, p. 106173, Aug. 2023, doi: 10.1016/J.ENGAPPAI.2023.106173.

[24] S. Kollem, C. Sirigiri, and S. Peddakrishna, "A novel hybrid deep CNN model for breast cancer classification using Lipschitz-based image augmentation and recursive feature elimination," Biomed Signal Process Control, vol. 95, p. 106406, Sep. 2024, doi: 10.1016/J.BSPC.2024.106406.

[25] X. Xu, Z. Liao, and Z. Xu, "Violent Physical Behavior Detection using 3D Spatio-Temporal Convolutional Neural Networks." [Online]. Available: www.ijacsa.thesai.org

[26] D. Sriveni and D. L. R, "An active learning driven deep spatio-textural acoustic feature ensemble assisted learning environment for violence detection in surveillance videos," Engineering Science and Technology, an International Journal, vol. 66, p. 102050, Jun. 2025, doi: 10.1016/J.JESTCH.2025.102050.

[27] J. Liu, P. Dai, G. Han, and N. Sun, "Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence," Computers and Electrical Engineering, vol. 106, p. 108614, Mar. 2023, doi: 10.1016/J.COMPELECENG.2023.108614.

[28] G. Sakthivinayagam, R. Easawarakumar, A. Arunachalam, and P. M, "Violence Detection System using Convolution Neural Network," International Journal of Electronics and Communication Engineering, vol. 6, no. 2, pp. 5–8, 2019, doi: 10.14445/23488549/ijece-v6i2p102.

[29] W. Ullah et al., "Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data," Future Generation Computer Systems, vol. 129, pp. 286–297, Apr. 2022, doi: 10.1016/j.future.2021.10.033.

[30] S. K. Parui, S. K. Biswas, S. Das, M. Chakraborty, and B. Purkayastha, "An Efficient Violence Detection System from Video Clips using ConvLSTM and Keyframe Extraction," 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, IEMECON 2023, no. June, pp. 1–5, 2023, doi: 10.1109/IEMECON56962.2023.10092302.

[31] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," Sensors, vol. 19, no. 11, pp. 1–15, 2019, doi: 10.3390/s19112472.

[32] N. Mumtaz et al., "An overview of violence detection techniques: current challenges and future directions," Artif Intell Rev, vol. 56, no. 5, pp. 4641–4666, 2023, doi: 10.1007/s10462-022-10285-3.

[33] R. Chatterjee and R. Halder, "Discrete Wavelet Transform for CNN-BiLSTM-based Violence Detection," in International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (ETAEERE -2020), 2020.

[34] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using Convolutional Neural Networks," Multimed Tools Appl, vol. 81, no. 26, pp. 38151–38173, Nov. 2022, doi: 10.1007/s11042-022-13169-4.

[35] Zanzan Lu, X. Xia, H. Wu, and C. Yang, "Violence Detection With Two-Stream Neural Network Based on C3D," International Journal of Cognitive Informatics and Natural Intelligence, vol. 15, no. 4, pp. 1–17, 2021, doi: 10.4018/IJCINI.287601.

[36] M. A. Soeleman, C. Supriyanto, D. P. Prabowo, and P. N. Andono, "Experimental Evaluation of Resampling Algorithms on the Imbalance Violence Video Detection," International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 260–268, 2022, doi: 10.14445/22315381/IJETT-V70I7P226.

[37] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," Neurocomputing, vol. 597, p. 128113, Sep. 2024, doi: 10.1016/j.neucom.2024.128113.

[38] M. Edo, V. Oubiña, and M. Svarc, "Machine learning and public policy: Early detection of physical violence against children," Child Youth Serv Rev, vol. 166, p. 107932, Nov. 2024, doi: 10.1016/j.childyouth.2024.107932.

[39] M. Bianculli et al., "A dataset for automatic violence detection in videos," Data Brief, vol. 33, p. 106587, 2020, doi: 10.1016/j.dib.2020.106587.

[40] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "A crowd analysis framework for detecting violence scenes," ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval, no. June, pp. 276–280, 2020, doi: 10.1145/3372278.3390725.

[41] M. Bianculli et al., "A dataset for automatic violence detection in videos," Data Brief, vol. 33, p. 106587, Dec. 2020, doi: 10.1016/J.DIB.2020.106587.

[42] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," IEEE Access, vol. 9, pp. 160580–160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[43] F. A. Memon, M. H. Memon, I. A. Halepoto, R. Memon, and A. R. Bhangwar, "Action Recognition in videos using VGG19 pre-trained based CNN-RNN Deep Learning Model," VFAST Transactions on Software Engineering, vol. 12, no. 1, pp. 46–57, Mar. 2024, doi: 10.21015/vtse.v12i1.1711.

[44] N. Honarjoo, A. Abdari, and A. Mansouri, "Violence Detection Using One-Dimensional Convolutional Networks," 2021 12th International Conference on Information and Knowledge Technology, IKT 2021, no. December, pp. 188–191, 2021, doi: 10.1109/IKT54664.2021.9685835.

[45] M. A. Soeleman, C. Supriyanto, D. P. Prabowo, and P. N. Andono, "Video Violence Detection Using LSTM and Transformer Networks Through Grid Search-Based Hyperparameters Optimization," International Journal of Safety and Security Engineering, vol. 12, no. 05, pp. 615–622, 2022, doi: 10.18280/ijsse.120510.

[46] T. Haque, F. F. Ahmed, S. M. I. Ahmed, and M. Siam, "Optical Flow based Violence Detection from Video Footage using Hybrid MobileNet and Bi-LSTM," no. September, 2023.

[47] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "State-of-the-art violence detection techniques in video surveillance security systems: A systematic review," PeerJ Comput Sci, vol. 8, pp. 1–41, 2022, doi: 10.7717/PEERJ-CS.920.

[48] N. Mumtaz, N. Ejaz, S. Aladhadh, S. Habib, and M. Y. Lee, "Deep Multi-Scale Features Fusion for Effective Violence Detection and Control Charts Visualization," Sensors, vol. 22, no. 23, pp. 1–15, 2022, doi: 10.3390/s22239383.

[49] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, Jan. 2020, doi: 10.1186/S12864-019-6413-7.

[50] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient Violence Detection in Surveillance," Sensors, vol. 22, no. 6, 2022, doi: 10.3390/s22062216.

[51] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Crowd Violence Detection from Video Footage," Proceedings - International Workshop on Content-Based Multimedia Indexing, vol. 2021-June, no. September, 2021, doi: 10.1109/CBMI50038.2021.9461921.

[52] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly Supervised Violence Detection in Surveillance Video," Sensors, vol. 22, no. 12, pp. 1–29, 2022, doi: 10.3390/s22124502.

[53] M. Biswas et al., "State-of-the-Art Violence Detection Techniques: A review," Asian Journal of Research in Computer Science, no. February, pp. 29–42, 2022, doi: 10.9734/ajrcos/2022/v13i130305.

[54] Md. B. Rahman, H. A. Mustafa, and M. D. Hossain, "Towards evaluating robustness of violence detection in videos using cross-domain transferability," Journal of Information Security and Applications, vol. 77, p. 103583, Sep. 2023, doi: 10.1016/j.jisa.2023.103583.

[55] S. A. Putri, "HU Variance Moment Optimizes Keyframe Selection Based on Deep Learning for Violence Detection," Journal of Applied Data Sciences, vol. 6, no. 2, pp. 1088–1101, May 2025, doi: 10.47738/jads.v6i2.648.